Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device

Article (Published Version)

Perception, 2011, volume 40, pages 1120 – 1135

doi:10.1068/p6952

Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device

David Brown§, Tom Macpherson, Jamie Ward¶#
School of Psychology, University of Sussex, Falmer, Brighton BN1 9QH, UK;
e-mail: jamiew@sussex.ac.uk; § and Research Centre in Psychology, Queen Mary, University of London, London, UK; # and Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK
Received 5 March 2011, in revised form 31 August 2011

Abstract. Sensory substitution devices convert live visual images into auditory signals, for example with a web camera (to record the images), a computer (to perform the conversion) and headphones (to listen to the sounds). In a series of three experiments, the performance of one such device ('The vOICe') was assessed under various conditions on blindfolded sighted participants. The main task that we used involved identifying and locating objects placed on a table by holding a webcam (like a flashlight) or wearing it on the head (like a miner's light). Identifying objects on a table was easier with a hand-held device, but locating the objects was easier with a head-mounted device. Brightness converted into loudness was less effective than the reverse contrast (dark being loud), suggesting that performance under these conditions (natural indoor lighting, novice users) is related more to the properties of the auditory signal (ie the amount of noise in it) than the cross-modal association between loudness and brightness. Individual differences in musical memory (detecting pitch changes in two sequences of notes) was related to the time taken to identify or recognise objects, but individual differences in self-reported vividness of visual imagery did not reliably predict performance across the experiments. In general, the results suggest that the auditory characteristics of the device may be more important for initial learning than visual associations.

1 Introduction

Sensory substitution devices convert one type of sensory signal into another. They may be used to facilitate the everyday functioning of individuals with some form of sensory loss, notably blindness. For example, there are devices that convert a visual image into touch (eg Bach-y-Rita 1972; Bach-y-Rita et al 1969; Bach-y-Rita and Kercel 2003) or sound (eg Arno et al 2001; Cronly-Dillon et al 1999, 2000; Hanneton et al 2010) thus enabling blind people to process visually derived information by one of their intact senses. One device that has attracted recent interest from researchers is called the vOICe (www.seeingwithsound.com Meijer 1992). The vOICe converts a visual image to an auditory signal (the capitalised middle letters spell out 'Oh, I see'). In this study we examine the effectiveness of various image-to-sound conversion algorithms and also user interactions with the device (whether it is mounted on the head or held in the hand) primarily using tasks of object recognition and localisation.

Visual-to-auditory devices provide a unimodal auditory input, but does the user process the signal in a purely auditory way or does the user transfer the auditory signal to routines specialised for visual or spatial processing? Blind users of the vOICe report detailed visual phenomenology (including lightness, shape, and depth) when wearing the device over a period of several months (Ward and Meijer 2010). Amedi et al (2007) trained sighted participants to use the device over several weeks (40 h of training) and examined their brain activity, using fMRI, when listening to the soundscapes relative to scrambled sounds or environmental sounds (eg 'moo'). The sighted expert users and two blind expert users activated a region of the lateral occipital cortex (LOC) that has been implicated in multisensory shape perception from sight or touch (non-expert users did not activate this region). This suggests that the experts were extracting shape

¶ Address all correspondence to Jamie Ward.

from the auditory signal using regions of the brain that do not normally respond to auditory stimuli. Auvray et al (2007) trained blindfolded sighted participants over 15 h using a hand-held webcam in a number of ecologically valid tasks, such as locating objects on a table or identifying objects. Interestingly, participants reported that using the device felt more like vision and less like hearing when attempting to locate objects rather than identify them. They interpreted their results in terms of sensorimotor theories of vision (eg O'Regan and Noe 2001). Namely, that hearing or touch will take on the phenomenological characteristics of vision when they have the same sensorimotor characteristics of vision. For example, if a visual object is occluded then it cannot be seen but occluded auditory objects can typically be heard. However, this kind of situation does apply to sounds that are converted from visual images, as in sensory substitution devices. These sounds obey the normal sensorimotor contingencies of visual objects and not auditory objects.

The hardware for a visual-to-auditory device itself consists of as little as a mobile phone or laptop computer, a webcam (to record the visual image), and the headphones (to listen to the signal). The software comprises an algorithm for converting visual features (brightness and spatial position) into auditory features (pitch, loudness, time, and stereo panning). The vOICe translates the visual image into a 2-D array of greyscale pixels. In the standard version, it then converts these pixels into auditory 'soundscapes' using the following three principles:

(i) Horizontal space → time. There is horizontal scanning of the image such that it is converted into sound from left-to-right over 1 s. Thus, an object on the left of space would be heard earlier in the soundscape. In stereo headphones there is also auditory panning from left to right.

(ii) Vertical space → pitch (frequency of pure tone). Each pixel in each column (ie time window) is associated with a pure tone. Higher pixels are associated with higher pitches.

(iii) Brightness → loudness (amplitude of pure tone). Brighter pixels are louder and black is silent.

This, however, is just one of many algorithms that could be used to convert an image into sound. The vOICe itself has a number of alternative settings. For example, reverse contrast flips the third rule above (so that bright is silent and dark is loud); a 'snail' setting displays the soundscape over 2 s (in the first rule above); and an edge-detection algorithm pre-processes the image before converting to sound so that edges are brighter (and, hence, louder). In addition, the user can vary how the device is used, such as holding the webcam in one's hand (like a flashlight) or wearing it on the head (like a miner's light). These four settings alone generate 16 possible user permutations and there are many more besides. At present, there is no published research that determines which settings are most beneficial and, hence, potential users of the device can be given no informed guidance on its optimal use.

Understanding the optimal settings for using a sensory substitution device is also important for developing a scientific understanding of how the brain learns to use such devices. Although the relative advantages of the different mappings have not been empirically explored with the device itself, they were guided by other research on cross-modal processing. For example, there is interference between pitch categorisation and the presence of irrelevant visual stimuli at different vertical positions (eg Evans and Treisman 2010). Similar effects have been reported in infants which suggests that this is not a culturally or linguistically based phenomenon (Wagner et al 1981). With regard to the association with loudness and brightness, participants explicitly judge loud stimuli to be brighter (Marks 1974, 1982). In more implicit tasks of association, sounds are judged to be louder if accompanied by an irrelevant visual stimulus (Odgaard et al 2004) and visual stimuli are judged to be brighter if accompanied by an irrelevant loud stimulus (Odgaard et al 2003).

The aims of this study are: to explore differences in the way that users interact with a sensory substitution device; to explore different algorithms for converting visual images to sounds; and to explore the impact of individual differences (eg in imagery, musical ability) on using the device. Auvray et al (2007) used a hand-held webcam because they believed that it would give the users more freedom to move the device. No study has directly compared performance with hand-held versus head-mounted devices, but we do so in experiments 1 and 3. In experiments 2 and 3, we compare different ways of converting images to sounds including normal and reversed-contrast. If users are relying on cross-modal pathways to reconvert the sound back into vision, then we predict that normal contrast may have an advantage because there is a general tendency to associate bright with loud and dark with quiet (eg Marks 2004), rather than vice versa. Finally, we consider how performance is modulated by individual differences in visual imagery (experiments 1 and 2) or in pitch perception and musical memory (experiments 2 and 3). Does performance by naive users depend on auditory abilities or visual-imagery abilities? Although anecdotal, Auvray et al (2007) noted that one of their best performing participants was a musician. The tests are performed on blindfolded sighted individuals for several reasons. Blind participants are a very heterogeneous group (early versus late onset, amount of residual vision) and a between-subjects design (with different individuals learning different settings) would require large groups. Also, whilst blind participants may be expected to perform better using these devices, they do not necessarily perform qualitatively differently (eg Amedi et al 2007; Sampaio et al 2001). Of course, it would be important to extend findings to this group in the future.

2 Experiment 1

This experiment contrasted different camera locations (hand-held versus head-mounted) in three different kinds of tasks that were motivated by the previous literature. The first task used simple 2-D geometric shapes. It was conceptually similar to studies by Kim and Zatorre (2008) and Cronly-Dillon and colleagues (Cronly-Dillon et al 1999, 2000). The second task was based on those used by Auvray et al (2007) but examined both object identification and localisation within the same task (to prevent the session from being too long). Auvray et al (2007) speculated that a hand-held device may allow "more extraction of the sensorimotor invariants while exploring the visual scene" but no study has directly contrasted this. The third task involved navigating past obstacles in a corridor and was loosely based on procedures used by Jones and Troscianko (2006) and Durette et al (2008). Correlations between individual differences in self-reported visual imagery and task performance were carried out.

2.1 *Method*

2.1.1 *Participants.* Eighteen participants took part, with equal numbers of males and females and an average age of 21 years (range 20 – 33 years). All reported normal or corrected-to-normal vision. They had no prior experience of the vOICe. The participants were randomly assigned to one of two groups ($N = 9$ in each): either wearing the head-mounted camera or holding the camera in their hands. These two sub-groups were matched in terms of mean age (age 21 years; head-mounted group: five males/four females; hand-held group: four males/five females). Participants were paid £5 per hour for their time.

2.1.2 *Overall procedure.* Before the experiment, the participants were told that they would be required to be blindfolded for up to 3 h and would have to learn to pick up objects, identify shapes, and navigate obstacles using sounds converted (via a computer) from images from a live camera. Previous studies suggest that it makes little difference whether participants are given formal instructions about the way the device converts

visual images to sounds or not (Kim and Zatorre 2008) and no such explanation of the device was given here. They filled-in a consent form and the Vividness of Visual Imagery Questionnaire (VVIQ—Marks 1973; for reliability of this measure see McKelvie 1995) before the tasks started. The tasks were always run in the same order: task 1 (shape discrimination/pre-training), then task 2 (locating and identifying objects), then task 3 (navigating obstacles), and finally task 1 again (shape discrimination/post-training). However, task 3 was introduced after an initial six participants had been tested, so the sample size for the tasks 1, 2, and 3 were eighteen, eighteen, and twelve, respectively (with equal numbers of head-mounted and hand-held cameras for each task). At the end of all the tasks, the blindfold and headphones were removed and the participants were debriefed.

2.1.3 *Task 1: Shape discrimination.* This task consists of a four-alternative forced choice identification of shapes (triangle, upright square, diamond, circle). The device was not worn for this task. Instead, soundscapes from sonified static images were used. The experiment was repeated both before sensorimotor use of the device (in tasks 2 and 3) and after.

Design. A 2×2 design was used contrasting location of camera used in training (head or hand) and before/after training as a within-subjects factor.

Materials. A large variety of stimuli were used to encourage participants to learn the general rules of the vOICe rather than rely on learned responses to a limited set of stimuli. All visual stimuli consisted of black shapes against a white background. The visual stimuli were converted into sound files prior to the start of the experiment using a feature of the vOICe software (normal contrast function, 1 s timing). There were four basic shapes. Each shape appeared in one of three vertical locations (centred at 25%, 50%, and 75% of height) which would manifest itself in the vOICe as shifts in the distribution of pitch. In addition, three different sizes were used (small, medium, and large) and the shapes were either solid or thick line-drawings. These different factors were fully counterbalanced: ie 4 shapes $\times$ 3 positions $\times$ 3 sizes $\times$ 2 shadings $=$ 72 unique stimuli. These stimuli were presented three times in random order, making a total of 216 trials in the task.

Procedure. Participants were seated at a table in front of a laptop and told that they would be listening to a number of sounds that had been converted from shapes with the vOICe. They were told that the vOICe converts visual images into sounds but were not explicitly informed of the rules of how this is achieved. One example of each shape's sound was played to them (large/centred/line-drawing) but the fact that there were many representations for each shape was not conveyed to the participants. They were then shown the four keys on the keyboard, and were then blindfolded with black-out. The participants put on a pair of overhead headphones, and when they were ready to begin they pushed any key on the keyboard and the sounds began. Stimuli were presented with EPrime 2.0. Each sound lasted for 1 s and repeated until the participant responded. Pilot testing revealed that participants find the task hard, and so accuracy was stressed to participants and timing of responses was not recorded. Participants were encouraged to guess if unsure. A beep after each response provided feedback to the participants telling them whether their response was correct or incorrect. There was a 1 s interval before the onset of the next sound. When the task had been completed the headphones were taken off but the blindfold was left on. The same procedure was conducted again at the end of the session (ie after tasks 2 and 3).

2.1.4 *Task 2: Locating and identifying objects.* This task is adapted from that reported by Auvray et al (2007) in which objects were placed on a table and the participant was required to either locate them or identify them. However, because of time constraints on participants we opted to combine these two tasks into a single experiment.

Design. The task used a 2×2 design contrasting the location of the camera between subjects (head-mounted versus hand-held) with the type of information within subjects (object location versus identity). The dependent measure was accuracy (%) for each type of information. The time taken (seconds) was also recorded for each trial.

Materials. The testing room contained a table with a pale wooden surface, plain carpet, white walls, and natural lighting in the form of a window behind the participants. The table contained a 3×3 grid of squares (ie 9 possible locations) with each individual square measuring 15 cm $\times$ 15 cm. The grid was visible to the experimenter but not the participant whilst using the device. There were four objects of similar size but different shape from each other. These consisted of an orange, a black computer mouse, a brown mug, and a black spatula. Each object was presented once in each location, giving a total of 36 unique trials. Participants had been shown the objects and the grid prior to task 1. The webcam was an MSI Star-Cam with a 56° field of view (and also for all subsequent tasks). It was either held in the hand or mounted on the top of blacked-out goggles.

Procedure. Participants had remained blindfolded from the first task and were told that they would be identifying both the location and the type of object. They were given up to 90 s to do this, although accuracy was stressed over speed. Participants were stood up and positioned at the end of the table with the grid nearest them. The device was switched on and a short amount of practice time was allowed (1 to 2 min) but not with the stimulus objects. The experimenter helped the participant to identify the edge of his/her 'visual' field by shining a torch light in the four corners of the camera's angle of view. The experimenter explained the set-up in terms of the position of the 3×3 grid (the participants reported being able to detect the table edge but could not see the grid) and the four objects used. Following this, the experiment began. During the experiment, participants were not allowed to touch the table or the object but they were otherwise free to explore. A tap on the shoulder indicated that the experimenter had started the stopwatch. A second tap after 60 s warned the participant that they had only 30 s to go, and three successive taps were used to indicate that the time was up and the participant must give an answer. If participants felt confident they had located/identified the object before the 90 s was up, they said "stop" and indicated their answers to the experimenter. Participants could name the position either by giving the location number (between 1 and 9) or by describing the position (eg top-right, bottom-middle, centre square, etc). After giving an answer, the participant was then free to touch the object and table in order to facilitate learning. The experimenter replaced the object and a tap to the shoulder indicated that the next trial should begin. After participants had completed the task they were allowed to take off the headphones, but told to leave the blindfold on for the next task.

2.1.5 *Task 3: Navigating obstacles.* In this task, participants had to navigate in a large indoor space by walking between four obstacles.

Design. There was one between-subjects factor: hand-held device versus head-mounted device. All participants were blindfolded. The within-subjects factor was trial number ($N = 8$). There were two dependent measures: time taken (seconds) and number of collisions with an obstacle.

Materials. There were four obstacles. Each obstacle was 1.4 m to 1.8 m in height and around 30 cm in depth and width. The obstacles were made out of cardboard boxes (two obstacles) and tall pot plants (two obstacles), and the size was chosen to be similar to people and small trees. The indoor space in which the obstacles were placed measured 11.58 m by 3.05 m. Each obstacle was arranged linearly down the centre and approximately equally spaced. There was natural lighting in the form of a window down one half of one of the longer walls.

Procedure. Participants were led, whilst blindfolded, out of the testing room used for tasks 1 and 2 and into the hallway. The experimenter then set up the obstacles (ie these were not seen prior to the task) whilst the participant was free to look around the room on the spot using the vOICe. The participants were then told that they had to walk to the other end of the corridor and touch the opposite wall. They were instructed to slalom between the four obstacles (either by an initial left or right turn, as they preferred) but without touching them. If the participants did make a collision, they were stopped and turned to face their final destination by the experimenter. A stopwatch was started at the beginning of each trial and ended when the participant touched the opposing wall. Upon reaching the opposite wall the participant was turned around and instructed to make the return journey whilst also being timed under the same conditions. Each participant navigated around the course on eight occasions (four in each direction). The task took approximately 40 min to complete. After this task the participants were allowed to remove their headphones, but not their blindfold. They were lead by the experimenter into the initial testing room where they completed task 1 again.

2.2 *Results and discussion*

Figure 1 summarises the results from task 1 in which participants had to identify the shape encoded by the soundscape. The participants were generally poor at this task but were nevertheless above chance in all four conditions (one-sampled t-tests, $p < 0.05$). A 2×2 ANOVA contrasting use of device with before/after training revealed significant improvements after training ($F_{1,16} = 207.51$, $p < 0.001$, $\eta^2 = 0.93$) but no effect of location of the device ($F_{1,16} = 0.11$, ns, $\eta^2 = 0.01$) and no interaction ($F_{1,16} = 0.83$, ns, $\eta^2 = 0.00$). It is to be noted that the device was not actually worn for this task (the sounds were pre-recorded from static images), although one could still have hypothesised that the intervening training involving different types of exploration could have affected performance differentially. The poor results obtained can be explained by the fact that body movements are indispensable to develop real perceptive abilities, and perhaps also the problem of identifying dark/silent objects against bright/loud backgrounds.

Figure 2 summarises the results from task 2 in which participants had to locate and identify objects. Considering accuracy first, a 2×2 ANOVA contrasting the different use of device (head versus hand) with location/identity revealed a significant main

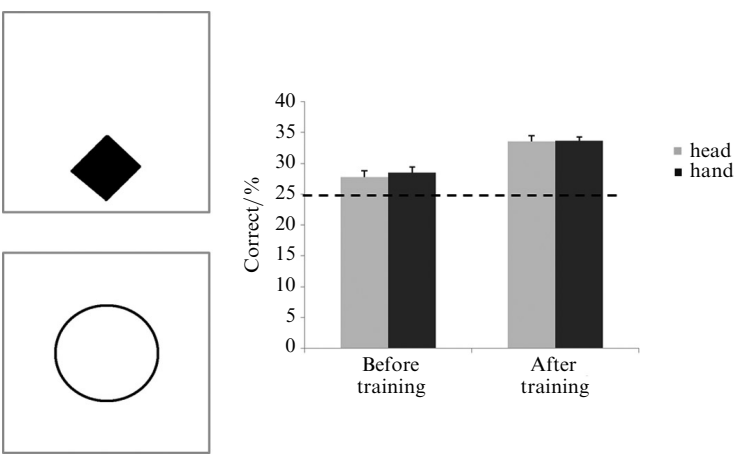

Figure 1. Accuracy at identifying simple geometric shapes (examples shown on the left), presented as sound files, before and after training with the vOICe (experiment 1, task 1). Error bars show one standard deviation; dashed line shows chance level of performance.

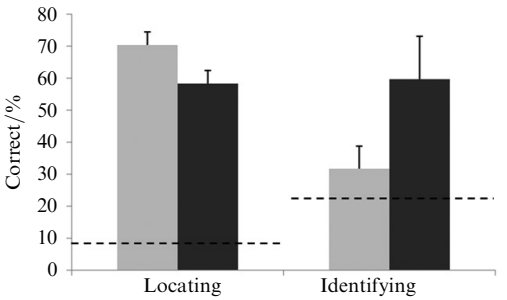

Figure 2. Accuracy at locating and identifying objects using two versions of the device in experiment 1 (task 2). Error bars show one standard deviation; dashed line shows chance level of performance.

effect of location/identity due to participants generally finding it easier to locate the object than identify it ($F_{1,16} = 111.13$, $p < 0.001$, $\eta^2 = 0.44$) and a significant main effect of device type ($F_{1,16} = 26.69$, $p < 0.001$, $\eta^2 = 0.63$) with the hand-held device performing better overall. There was also a significant interaction between these factors ($F_{1,16} = 126.31$, $p < 0.001$, $\eta^2 = 0.50$). This was because the head-mounted device was significantly better at locating the position of the object than the hand-held device ($t_{16} = 6.07$, $p < 0.001$). However, the reverse was true for identifying the object. The hand-held device was significantly better at identifying the object than the head-mounted device ($t_{16} = 10.22$, $p < 0.001$).

The participants' accuracy did not significantly improve over the 36 trials (correlations between percentage correct and trial number for each device and each type of information all non-significant). However, they did get significantly faster with practice reducing their time from 90 s down to 49 s (head-mounted: $r = 0.98$, $p < 0.001$; hand-held: $r = 0.99$, $p < 0.001$). A 2×36 ANOVA contrasting location of camera with trial number revealed a main effect of device ($F_{1,16} = 11.25$, $p < 0.001$) such that the head-mounted device had a small (3.5 s overall) but reliable benefit. This did not interact with trial number.

For task 3, in which participants had to navigate a route around four obstacles, the average time taken was 323 s, and the average improvement over the eight trials was 80 s. A 2×8 ANOVA comparing location of camera (head versus hand) versus trial number (1 to 8), with the dependent variable being time taken, revealed a significant main effect of trial number ($F_{7,70} = 33.21$, $p < 0.001$, $\eta^2 = 0.75$) but no main effect of camera location ($F_{1,10} = 0.19$, $\eta^2 = 0.02$) and no interaction ($F_{7,70} = 0.89$, $\eta^2 = 0.02$). The same pattern was found when number of collisions was entered as a dependent variable [main effect of trial number ($F_{7,70} = 14.62$, $p < 0.001$, $\eta^2 = 0.58$), main effect of location ($F_{1,10} = 0.91$, $\eta^2 = 0.08$), interaction ($F_{7,70} = 0.62$, $\eta^2 = 0.02$)].

The head-mounted group and hand-held group did not differ from each other in terms of self-reported vividness of visual imagery assessed prior to the study ($t_{16} = 0.14$, ns). However, there is evidence that this factor makes an independent contribution to their ability to use the device. In task 2, when VVIQ ratings are entered as a covariate in the 2×2 ANOVA (with dependent variable being accuracy) the effect of this variable is significant ($F_{1,15} = 6.39$, $p < 0.05$, $\eta^2 = 0.11$) but it did not interact with the other variables. VVIQ did not affect the timing in task 2 or task 3, or accuracy in task 1 when entered as a covariate in the appropriate ANOVAs.

In summary, the results of these three tasks demonstrate that this form of sensory-substitution device can be effective in a range of situations with no explicit training. Identifying sonified images of geometric shapes produced the least impressive results (around 33% correct with a chance level of 25%). However, performance is generally better when the participant is able to wear or carry the device and performance is assessed using objects placed on a table (task 2). This is consistent with the notion that sensory-motor interactions are important for successfully interpreting the signal.

For instance, by providing multiple views of an object and noticing how the soundscape changes when the device is moved laterally or back-and-forth may provide cues to its size, shape, and location. Wearing the device on the head arguably affords less sensory-motor interactions than holding it in the hands (eg Auvray et al 2007) and the findings from task 2 are broadly consistent with this. However, the effects may tend to be task-specific. The head-mounted device tended to be better for locating objects perhaps because it tended to be held further away and with less sudden movements. Of course, the actual location of the device is only important insofar as the location affords different kinds of changes to the visual image/auditory signal. Given that we did not instruct them in this regard, participants could have adopted a hand-like strategy with their head and a head-like strategy with their hand (and we would predict no difference in that scenario). We did not seek to control this, because we were interested in naturalistic use of the device when placed in differing locations. Interestingly, there was an effect of visual imagery, but the effects were limited to one measure on one task (accuracy on the identification/localisation task). There was no difference between the devices in the navigation task.

3 Experiment 2

The aim of this experiment was to investigate other parameters of the vOICe aside from location of the device on the body (which was always head-mounted). Two different image-to-sound conversion algorithms were compared in a between-subjects design. First, the brightness – loudness mapping was either normal contrast (ie light is loud, dark is quiet) or reversed-contrast (ie light is quiet, dark is loud). Based on the literature of cross-modal processing (eg Marks 1974, 2004) we expected the former to outperform the latter. Second, we contrasted a normal (1 s) scanning time with a slower, half-speed setting. It was predicted that participants would be more accurate at recognising objects when the auditory signature is longer, although this would not necessarily improve localisation given that there would be a longer lag between moving the device and hearing the change in the auditory signal. Finally, in addition to examining individual differences in vividness of visual imagery (as in experiment 1) we also examined individual differences in two measures of auditory perception.

3.1 *Method*

3.1.1 *Participants.* Thirty-two participants took part in the main tasks, with an average age of 21 years (range 20 – 33 years). All reported normal or corrected-to-normal vision. They had no prior experience of the vOICe. The participants were pseudo-randomly assigned to one of four groups ($N = 8$ in each), with the constraint that each group had five females and three males. A posteriori analyses revealed that the ages of the groups did not differ ($F_{3,28} = 1.26$, ns). The four groups were formed by orthogonally contrasting two settings of the device: contrast [normal (bright = loud) versus reversed (dark = loud)], and scanning time (1 s versus 2 s).

3.1.2 *Procedure.* Before taking part in the main tests, participants were asked to complete the VVIQ (Marks 1973). In addition, they were given two tests of auditory perception developed at the Beth Israel, Harvard Medical School (www.tonometric.com). Participants wore headphones for the test and the volume was adjusted for comfort. The musical memory test consisted of 36 trials involving two musical phrases played successively that were either the same or had a tonal change. The test gives a percentage-correct score. The adaptive pitch test plays a series of two short tones and asks if the second is higher or lower in pitch than the first. The interval between tones gradually decreases but the tones are never identical. The result, expressed in Hertz, is the difference between two tones that can be reliably discriminated at 500 Hz (a smaller difference indicates better pitch perception).

The procedure for experiment 2 was very similar to that in experiment 1, but consisted of only two tasks: namely the object identification and location task, and the navigation task. The task of detecting shapes was omitted because performance in experiment 1 was poor. All participants wore the head-mounted device. Unlike in experiment 1, the order of the two tasks was fully counterbalanced across participants. There were also some minor changes within each task. For the object location and identification task, the same room and the same objects were used as before. However, the participant faced a different edge of the table (the left and far edges were bounded by a wall). For the navigation task, the same room and obstacles were used as before but the positioning of the obstacles was jittered by around 30 cm after each attempt to prevent participants from relying on their memory of the position of the obstacles.

3.2 *Results and discussion*

The three tests of individual differences in visual imagery and musical ability all showed a wide variety of scores: vividness of visual imagery (mean = 51.4, SD = 11.7, range = 22–77); pitch discrimination (mean = 6.81, SD = 6.11, range = 0.68–26.4); and musical memory (mean = 74.7, SD = 8.7, range = 52.8–86.6). Individual scores were entered as covariates on the various sensory substitution tasks.

The results of the object recognition and localisation task are summarised in figure 3. The results were analysed with a $2 \times 2 \times 2$ mixed ANOVA comparing contrast (between participants, normal versus reversed), scanning speed (between participants, normal versus half-speed), and task (within participants, identify versus locate object). The dependent variable is the percentage correct. There was a significant main effect of contrast ($F_{1,25} = 10.09$, $p < 0.005$, $\eta^2 = 0.28$), with the reverse-contrast condition outperforming the normal-contrast condition. This is the opposite finding to that predicted from the literature on cross-modal audio-visual perception. The effect of contrast interacted with task ($F_{1,25} = 4.60$, $p < 0.05$, $\eta^2 = 0.14$). That is, the effect of contrast was greater for identifying objects (62.2% versus 45.7% for reversed-contrast and normal-contrast, respectively) than it was for locating them (37.0% and 29.3%, respectively). There were no main effects of task ($F_{1,25} = 1.81$, $\eta^2 = 0.05$) or scanning speed ($F_{1,25} = 0.52$, $\eta^2 = 0.01$). No other interactions approached significance and none of the covariates of vividness of visual imagery, pitch perception, or musical memory had an effect on recognition accuracy (all ps > 0.10).

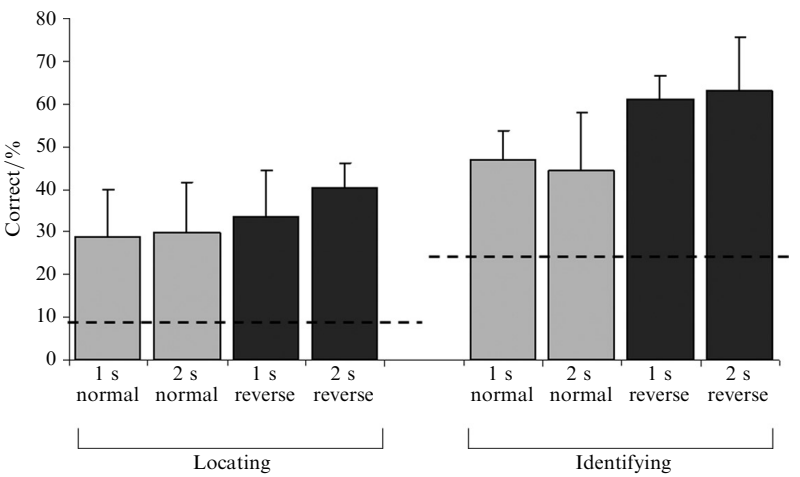

Figure 3. Accuracy at locating and identifying objects depending on different contrast settings (normal versus reversed) and scanning speeds (normal, 1 s versus slow, 2 s). Error bars show one standard deviation; dashed line shows chance level of performance.

The participants' accuracy did not significantly improve over the 36 trials (identification: $r = 0.05$; localisation: $r = 0.21$). However, they did get significantly faster with practice ($r = 0.88$, $p < 0.001$). The time taken to complete the task was entered as a dependent variable in a $2 \times 2 \times 36$ ANOVA comparing contrast (reverse versus normal), scanning time (1 s versus 2 s), and trial number (1 to 36). There was no main effect of contrast setting ($F_{1,25} = 0.80$, $\eta^2 = 0.03$), no main effect of scanning time ($F_{1,25} = 0.75$, $\eta^2 = 0.03$), and no interaction between them ($F_{1,25} = 0.31$, $\eta^2 = 0.01$). As such, the increased accuracy in the reversed-contrast condition is not achieved through slower performance. There was a main effect of trial number ($F_{35,875} = 1.78$, $p < 0.005$, $\eta^2 = 0.05$), indicating that participants became faster with practice. None of the covariates of vividness of visual imagery, pitch perception, or musical memory had an effect on search time (all ps > 0.10). However, there was a significant interaction between musical memory and trial number ($F_{35,875} = 1.96$, $p < 0.001$, $\eta^2 = 0.06$). This was due to the fact that those with better musical memory tended to get faster with practice than those with worse musical memory. This is illustrated in figure 4. The participants' accuracy did not significantly improve over the 36 trials (correlations between accuracy at locating or recognising objects and trial number are non-significant).

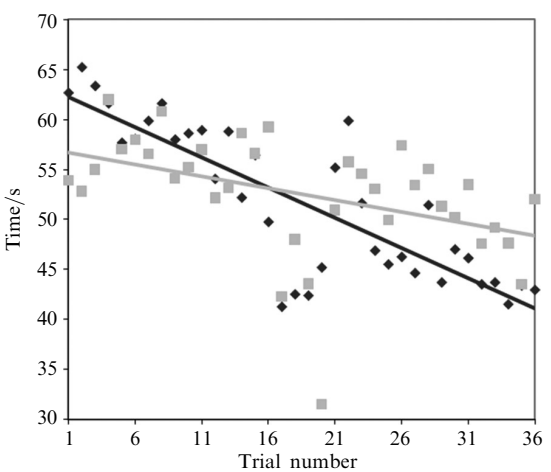

Figure 4. Time taken to locate/identify objects in experiment 2 as a result of practice over the 36 trials. The darker and lighter squares represent participants with above average and below average musical memory, respectively, taking a median split of participants.

The results from the navigation task were entered as a $2 \times 2 \times 8$ ANOVA comparing contrast (normal versus reversed), scanning time (1 s versus 2 s), and trial number (1 to 8). The dependent variable was time taken, in seconds. None of the main effects or covariates reached significance (all ps > 0.10). In experiment 1 participants showed evidence of improvement on this task (ie a main effect of trial number) whereas in experiment 2 they did not. However, in experiment 1 participants may have memorised the exact location of the obstacles, and could have improved by relying less on the vOICe and more on memory. In experiment 2, this effect was reduced by jittering the location of the objects from trial to trial. More importantly for the present discussion, none of the critical changes to the device settings affected performance during the navigation task.

In summary, the results of this experiment suggest that a reverse-contrast setting (bright = quiet) outperformed a normal-contrast setting, particularly for identifying objects. There was no evidence that individual differences in visual imagery or pitch perception affected performance, but there was some evidence that musical memory affected the learning rate. Collectively, the results are more consistent with the notion that, with minimal exposure (< 2 h), performance with this sensory substitution device depends more on auditory processes than cross-modal correspondences. Rooms with natural lighting may tend to generate soundscapes with multiple loud components

(corresponding to large reflective surfaces such as walls and windows), and these are likely to be reduced with a reverse-contrast setting. It would be interesting to know whether normal-contrast setting would outperform reverse-contrast setting if the auditory features of the corresponding soundscapes were equated. Participants are able to use pitch and temporal cues to locate/recognise objects but this could be (initially) via auditory – motor or auditory – spatial associations rather than auditory – visual. Although presenting information over a longer time window (2 s) was expected to improve performance, it also leads to sluggishness in updating the soundscape as a result of participant movements. This may explain why no difference was found between the faster and slower settings.

4 Experiment 3

Experiment 3 was designed to replicate and extend previous findings. The navigation task was dropped and, unlike in previous experiments, the object localisation/identification task was separated into two different tasks using either nine objects in a single location (for the identification task) or one object in nine possible locations (in the localisation task). The location of the device was varied between participants as either hand-held or head-mounted (as in experiment 1) and two different image-to-sound algorithms were contrasted (between subjects). The best performing algorithm from experiment 2 (reverse-contrast) was compared with the inbuilt edge enhancement setting of the vOICe. The edge enhancement setting filters the image using a Sobel operator to detect changes in luminance. Edges then appear bright (loud), but the vOICe then also blends in a low intensity (quieter) version of the original image to keep some of the original surface shading. Some examples of the experimental set-up shown in normal-contrast, reverse-contrast, and edge-enhancement (of normal-contrast) are shown in figure 5.

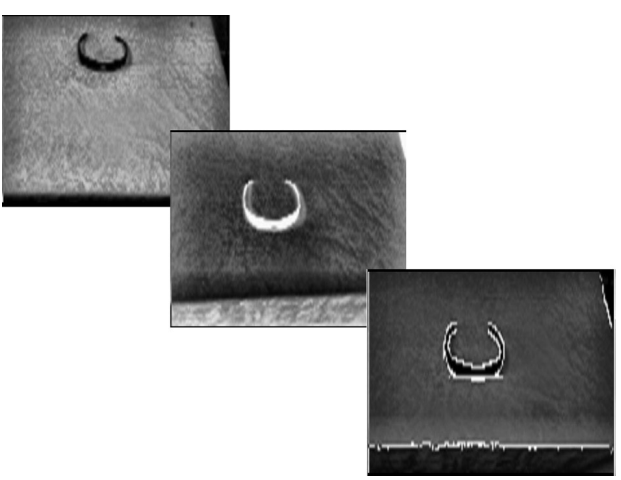

Figure 5. Examples of different settings of the vOICe. From left to right: normal-contrast, reverse-contrast, and edge-enhancement. Participants would not see these images but would hear the corresponding 'soundscape'.

4.1 *Method*

4.1.1 *Participants.* Twenty-four participants (twelve male, twelve female) took part in both tasks with an average age of 24 years (range = 19 – 36). All participants reported normal or corrected-to-normal vision and hearing. None had taken part in previous studies of the vOICe. Assignment to one of four experimental conditions (hand/head × reverse-contrast/edge-enhancement) was pseudo-random, with the constraint that there were equal numbers of males and females in each group. Participants were remunerated with course credits.

4.1.2 *Procedure.* The order of the two tasks was counterbalanced across participants. The web camera, goggles, computer, and headphones were the same as in experiment 2. The scanning time was 1 s throughout.

Localisation task. The table was covered with a white sheet with a 3×3 matrix (45 cm $\times$ 45 cm in total) marked in pencil on the surface. This was visible to the experimenter but not to the participant using the vOICe. The target object was a small dark-grey 'cat' toy. Participants were informed that they would be required to locate (in reference to the grid) the test object. Though each trial was timed, it was emphasised that accuracy was of primary importance. The vOICe was fitted and a comfortable volume level set. The grid dimensions within the vOICe's visual field were demonstrated by placing an object in the four grid extremities and centre square, and the participants told to locate the objects. The objects were then removed and the task started. Each trial commenced with the webcam being obscured whilst the object was placed in the test area. A tap on the participant's shoulder indicated the trial had started. A second tap indicated 60 s had elapsed, whilst three taps at 90 s informed the participant that they were required to give an answer. Location could be described by either the numbers 1–9 or by description (top-right, bottom-middle, centre-square, etc). Although 90 s were allocated for each trial, participants were free to give an answer at any time. To facilitate learning, feedback was given after every trial by moving the participant's hand to the chosen location then, if necessary, to the correct location for tactile confirmation. The object was presented four times in each square, giving a total of 36 trials. Different trial orders were used for the different between-subjects conditions.

Identification task. For this task 9 objects were used: spatula, TV remote control, USB cable (coiled), black CD case, clear CD case (edges blacked out), infant's windup toy, mug, large comb, sunglasses (open). All were dark in colour. There was no training on the objects with the vOICe prior to the task. There was, however, a tactile exploration and verbal description of the objects before the task. The general procedure was the same as that in the localisation task (including tactile feedback) except that the test object was always placed in the same grid location. Each of the nine objects was presented four times giving a total of 36 trials.

4.2 *Results and discussion*

Figure 6 summarises the results. For both tasks, chance level of performance is 11% and in all conditions performance with the device exceeds chance (confirmed with one-sample *t*-tests, $p < 0.005$ in all cases). A $2 \times 2 \times 2$ mixed ANOVA was conducted with between-subjects factors of sensor location (head versus hand) and image-to-sound

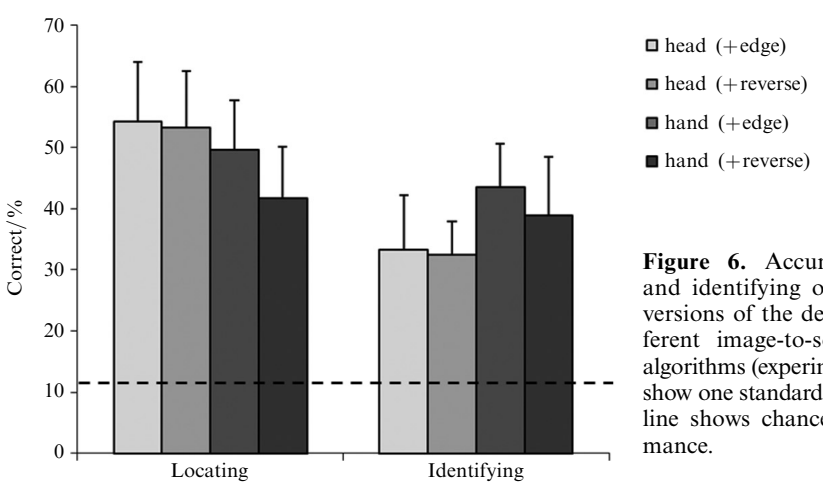

Figure 6. Accuracy at locating and identifying objects using two versions of the device and two different image-to-sound conversion algorithms (experiment 3). Error bars show one standard deviation; dashed line shows chance level of performance.

algorithm (reverse-contrast versus edge-enhancement) and task as a within-subject factor (identification versus localisation). The dependent variable was percentage correct. There was no main effect of algorithm ($F_{1,20} = 1.54$, $\eta^2 = 0.07$) and this factor did not interact with any other condition. There was, however, a main effect of task ($F_{1,20} = 45.87$, $p < 0.001$, $\eta^2 = 0.54$), with localisation being easier overall. The effect of task interacted with sensor location ($F_{1,20} = 19.46$, $p < 0.001$, $\eta^2 = 0.23$). This was because the head-mounted device was significantly better at locating the position of the object relative to the hand-held device ($t_{22} = 2.21$, $p < 0.05$), but the hand-held device was significantly better at identifying the objects relative to the head-mounted device ($t_{22} = 2.63$, $p < 0.05$). There were no other significant effects or interactions. Adding musical memory as a covariate does not affect the pattern described above, and this makes no independent contribution. Musical memory does, however, make a difference when one considers response time (see below).

The participants' accuracy did not significantly improve over the 36 trials (identification: $r = -0.11$; localisation: $r = -0.21$). However, they did get significantly faster with practice (identification: $r = 0.91$, $p < 0.001$; localisation: $r = 0.93$, $p < 0.001$). Two separate $2 \times 2 \times 36$ ANOVAs were run taking into account sensor location (head versus hand), image-to-sound conversion algorithm (edge-enhanced versus reversed), and trial number (36 levels) on the localisation task and identification task separately. Performance on the musical memory task was entered as a covariate. The results were generally similar across the two tasks. Participants get faster with practice (main effect of trial number; localisation task $F_{35,665} = 2.29$, $p < 0.001$, $\eta^2 = 0.09$; identification task $F_{35,665} = 1.71$, $p < 0.001$, $\eta^2 = 0.07$). Performance on the musical memory task had no main effect in its own right (effect of covariate; localisation task $F_{1,19} = 0.63$, $\eta^2 = 0.03$; identification task $F_{1,19} = 0.50$, $\eta^2 = 0.02$), but on both tasks this interacted significantly with trial number (localisation task $F_{35,665} = 1.79$, $p < 0.005$, $\eta^2 = 0.07$; identification task $F_{35,665} = 1.45$, $p < 0.05$, $\eta^2 = 0.06$). This replicates the finding of experiment 2, although the pattern is somewhat different: those with better musical memory perform faster on earlier trials but these advantages diminish as the experiment progresses.

In summary, experiment 3 replicates the basic finding of experiment 1 by showing an interaction between position of the camera and task (localisation and identification), in a modified experimental design in which the two tasks are performed separately and equated for chance level of performance. Edge-enhancement (ie making edges louder) is comparable to reverse contrast (bright = quiet, dark = loud) under these experimental conditions. Both manipulations tend to reduce the amount of auditory noise (in a light room), but it would be interesting for future research to determine how these two factors combine (ie edge-enhancement applied to reverse-contrast images).

5 General discussion

What is the optimal configuration for translating visual information into an auditory signal in a sensory substitution device, bearing in mind the constraints of the users' sensory, motor, and learning mechanisms? This study aimed to address this question using a visual-to-auditory sensory substitution device, the vOICe. In addition to considering changes to the image-to-sound conversion algorithm (scanning time, contrast settings), we also considered individual differences in the user (in terms of visual imagery and auditory perception), as well as different ways of interacting with the device (on the head or held in the hand). The results can be summarised as follows:

- A head-mounted device is better for locating objects than a hand-held device, whereas the opposite is true for identifying an object. That is, the optimal way of interacting with the device is task-specific.

- Reverse-contrast (ie dark = loud, bright = quiet) outperformed normal-contrast, but the effect was greater for recognising objects than locating them spatially. Edge-enhancement (making edges louder) was equivalent to reverse-contrast.
- Differences in the temporal length of the soundscape (1 s versus 2 s) had no effect on any of the tasks used here.
- There was evidence that participants with better musical memory (ie ability to compare musical phrases containing a tonal difference) show different learning rates. Individual differences in pitch discrimination and visual imagery had no reliable effects.

With regards to the difference between a hand-held versus head-mounted device, Auvray et al (2007) had speculated that the former might outperform the latter. However, there are other considerations that might be important. Proulx et al (2008) argued that a hand-held device may impede free use of the hands. Blind users tend to use head or body mounted set-ups (Ward and Meijer 2010). It was observed that, when using the hand-held device, participants tended to scan the table until the device was above the object. This may enable the object to be more easily identified and it may be possible to locate the object using proprioceptive knowledge of the position of the hand-held device (which we did not test), although not necessarily using the system based here which is defined relative to the table. A head-mounted device may tend to afford a 'bigger picture' enabling the object to be located relative to the table edges, rather than relative to the body. These results are unlikely to be due to the head or hand per se but rather how these parts of the body can be moved and positioned to elicit changes in the auditory signal. In principle, the body part itself versus its mode of exploration could be studied independently (eg by moving the head in a hand-like way and moving the hand in a head-like way). In addition, it is important that future research develops paradigms for assessing localisation of objects with sensory substitution devices using measures of reaching and grasping (eg as in Proulx et al 2008) which may depend less on consciously reportable knowledge of position (eg Goodale and Milner 1992).

With regards to the image-to-sound conversion algorithm used, our findings suggest that the best performing algorithms are not necessarily those that are derived from research on cross-modal auditory–visual correspondences. Instead, the most efficient algorithms tend to reduce the amount of auditory noise by reducing the number of pixels that are bright/loud. In general, this supports the conclusion that brief (under 2 h) interactions with a visual-to-auditory sensory substitution device are more likely to be affected by processes such as auditory scene analysis rather than attempts to process the signal visuo-spatially. In support of this, individual differences in a simple test of musical memory reliably affected performance (in terms of speed of responding), whereas individual differences in visual imagery were not found reliably (they were limited to one measure on one task, and this was not subsequently replicated). Poirier et al (2007) argued that visual imagery may be important for detecting patterns conveyed in a similar visual-to-auditory device. They based this conclusion on increased activity (measured with fMRI) in visual regions after a short amount (2 h) of training with their device. However, given the nature of fMRI it is unclear whether participants were attempting to use visual imagery or whether visual imagery actually supported task performance. It is also to be noted that their stimuli were far simpler than the ones used here (with the exception of task 1 in experiment 1) which may require less attention to the auditory signal. Given that blind-expert users of the device report visuo-spatial phenomenology (Ward and Meijer 2010) and functional imaging and TMS suggests that expert users of this device recruit regions involved in shape processing not normally activated by auditory perception (Amedi et al 2007; Merabet et al 2009), it remains unclear how and when this hypothesised transition from audition to auditory–visual processing occurs but it is likely to take several days rather than several hours of training, at least in ecologically valid settings.

It is possible that a normal bright-to-loud algorithm may have benefits when the device is used for sensory augmentation rather than sensory substitution, ie in those conditions in which a sensory substitution device accompanies a (perhaps degraded) visual input. For example, some blind users of the device have some residual light perception. Under these conditions, the normal-contrast setting, which was found not to be optimal in this study, may fare better than reverse-contrast. This could be tested in sighted participants using goggles that degrade acuity rather than the more conventional method of blindfolding which prevents vision altogether. It would also be interesting to explore the use of these devices in patients with visual field defects (eg hemianopia) who have normal acuity but spatially limited vision.

In summary, the results of this study suggest that blindfolded sighted participants are able to use a visual-to-auditory sensory substitution device in naturalistic experiments to locate and identify objects with above chance accuracy. This occurs with no explicit instructions about how the device works, but emerges through exploring how the auditory signal changes for different kinds of objects and different kinds of interactions (eg by moving the device from side to side). We have identified some conditions in which performance is optimised but it remains for further research to establish how general these findings are.

References

Amedi A, Stern W, Camprodon J A, Bermpohl F, Merabet L, Rotman S, Hemond C, Meijer P, Pascual-Leone A, 2007 "Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex" *Nature Neuroscience* **10** 687 – 689

Arno P, De Volder A G, Vanlierde A, Wanet-Defalque M-C, Streel E, Robert A, Sanabria-Bohorquez S, Veraart C, 2001 "Occipital activation by pattern recognition in the early blind using auditory substitution for vision" *NeuroImage* **13** 632 – 645

Auvray M, Hanneton S, O'Regan J K, 2007 "Learning to perceive with a visual-auditory substitution system: Localisation and object recognition with 'The vOICe' " *Perception* **36** 416 – 430

Bach-y-Rita P, 1972 *Brain Mechanisms in Sensory Substitution* (New York: Academic Press)

Bach-y-Rita P, Collins C C, Saunders F A, White B, Scadden L, 1969 "Vision substitution by tactile image projection" *Nature* **221** 963

Bach-y-Rita P, Kercel S W, 2003 "Sensory substitution and the human – machine interface" *Trends in Cognitive Sciences* **7** 541 – 546

Cronly-Dillon J, Persaud K, Gregory R P F, 1999 "The perception of visual images encoded in musical form: A study in cross-modality information transfer" *Proceedings of the Royal Society of London, Series B* **266** 2427 – 2433

Cronly-Dillon J, Persaud K C, Blore R, 2000 "Blind subjects construct conscious mental images of visual scenes encoded in musical form" *Proceedings of the Royal Society of London, Series B* **267** 2231 – 2238

Durette B, Louveton N, Alleyson D, Herault J, 2008 "Visuo – auditory sensory substitution for mobility testing: The VIBE" paper presented at the 10th European Conference on Computer Vision, Marseille

Evans K K, Treisman K, 2010 "Natural cross-modal mappings between visual and auditory features" *Journal of Vision* **10**(1) 1 – 12

Goodale M A, Milner A D, 1992 "Separate visual pathways for perception and action" *Trends in Neurosciences* **15** 20 – 25

Hanneton S, Auvray M, Durette B, 2010 "The Vibe: a versatile vision-to-audition sensory substitution device" *Applied Bionics and Biomechanics* **7** 269 – 276

Jones T, Troscianko T, 2006 "Mobility performance of low-vision adults using an electronic mobility aid" *Clinical and Experimental Optometry* **89** 10 – 17

Kim J-K, Zatorre R J, 2008 "Generalized learning of visual-to-auditory substitution in sighted individuals" *Brain Research* **242** 263 – 275

McKelvie S J, 1995 "The VVIQ and beyond: Vividness and its measurement" *Journal of Mental Imagery* **19** 197 – 252

Marks D F, 1973 "Visual imagery differences in the recall of pictures" *British Journal of Psychology* **64** 17 – 24

Marks L E, 1974 "On associations of light and sound: The mediation of brightness, pitch, and loudness" *American Journal of Psychology* **87** 173 – 188

Marks L E, 1982 "Bright sneezes and dark coughs, loud sunlight and soft moonlight" *Journal of Experimental Psychology: Human Perception and Performance* **8** 177 – 193
Marks L E, 2004 "Cross-modal interactions in speeded classification", in *The Handbook of Multisensory Processes* Eds G Calvert, C Spence, B E Stein (Cambridge, MA: MIT Press)
Meijer P B L, 1992 "An experimental system for auditory image representations" *IEEE Transactions on Biomedical Engineering* **39** 112 – 121
Merabet L B, Battelli L, Obretenova S, Maguire S, Meijer P, Pascual-Leone A, 2009 "Functional recruitment of visual cortex for sound encoded object identification in the blind" *NeuroReport* **20** 132 – 138
Odgaard E C, Arieh Y, Marks L E, 2003 "Cross-modal enhancement of perceived brightness: Sensory interaction versus response bias" *Perception & Psychophysics* **65** 123 – 132
Odgaard E C, Arieh Y, Marks L E, 2004 "Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation" *Cognitive, Affective and Behavioral Neuroscience* **4** 127 – 132
O'Regan J K, Noe A, 2001 "A sensorimotor account of vision and visual consciousness" *Behavioral and Brain Sciences* **24** 939 – 1031
Poirier C, De Volder A G, Tranduy D, Scheiber C, 2007 "Pattern recognition using a device substituting audition for vision in blindfolded sighted subjects" *Neuropsychologia* **45** 1108 – 1121
Proulx M J, Stoerig P, Ludowig E, Knoll I, 2008 "Seeing 'where' through the ears: Effects of learning by doing and long-term sensory deprivation on localization based on image to sound conversion" *PLoS One* **3** e1840
Sampaio E, Maris S, Bach-y-Rita P, 2001 "Brain plasticity: 'visual' acuity of blind persons via the tongue" *Brain Research* **908** 204 – 207
Wagner S, Winner E, Cicchetti D, Gardner H, 1981 "'Metaphorical' mapping in human infants" *Child Development* **52** 728 – 731
Ward J, Meijer P, 2010 "Visual experiences in the blind induced by an auditory sensory substitution device" *Consciousness and Cognition* **19** 492 – 500

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 40 2011

www.perceptionweb.com